Journal of Machine Learning Research () Submitted ; Published

A Consistent Information Criterion for Support Vector Machines in Diverging Model Spaces

Xiang Zhang XZHANG23@NCSU.EDU
Yichao Wu WU@STAT.NCSU.EDU
Department of Statistics
North Carolina State University
Raleigh, NC 27695, USA

Lan Wang WANGX346@UMN.EDU
Department of Statistics
The University of Minnesota
Minneapolis, MN 55455, USA

Runze Li RZLI@PSU.EDU
Department of Statistics and The Methodology Center
The Pennsylvania State University
University Park, PA 16802-2111, USA

Editor:

Abstract

Information criteria have been popularly used in model selection and proved to possess nice theoretical properties. For classification, Claeskens et al. (2008) proposed support vector machine information criterion for feature selection and provided encouraging numerical evidence. Yet no theoretical justification was given there. This work aims to fill the gap and to provide some theoretical justifications for support vector machine information criterion in both fixed and diverging model spaces. We first derive a uniform convergence rate for the support vector machine solution and then show that a modification of the support vector machine information criterion achieves model selection consistency even when the number of features diverges at an exponential rate of the sample size. This consistency result can be further applied to selecting the optimal tuning parameter for various penalized support vector machine methods. Finite-sample performance of the proposed information criterion is investigated using Monte Carlo studies and one real-world gene selection problem.

Keywords: Bayesian Information Criterion, Diverging Model Spaces, Feature Selection, Support Vector Machines

1. Introduction

We consider binary classification using linear support vector machines (SVMs). It is well known that the standard SVM uses all features while constructing the classification rule. In the extreme case of regression that the number of features is much larger than the sample size, if the true model is non-sparse, no method can identify the truth correctly due to the limited information from the data. This is the so-called *curse of dimensionality*. In many important applications, however, it is reasonable to simplify the problem by assuming the true model to be sparse. For example, in cancer classification using genomic data where the

number of probes (or genes) can be tens of thousands and the number of patient samples is typically only a few dozens, biologists find it plausible to assume that only a small subset of genes are relevant. In this case, it is more desirable to build a classifier based only on those relevant genes. Yet in practice, it is largely unknown which genes are relevant and this calls for feature selection methods. The potential benefits of feature selection include reduced computational burden, improved prediction performance and simple model interpretation. See Guyon and Elisseeff (2003) for more discussions. Our goal in this paper is consistent feature selection for the SVM.

There has been a rich literature on feature selection for the SVM. Weston et al. (2000) proposed a scaling method to select important features. Guyon et al. (2002) suggested the SVM recursive feature elimination (SVM RFE) procedure. It has been shown that the SVM can be fitted in the regularization framework using the hinge loss and the L_2 penalty (Wahba et al., 1999). Thereafter various forms of penalized SVMs have been developed for simultaneous parameter estimation and feature selection. Bradley and Mangasarian (1998), Zhu et al. (2004) and Wegkamp and Yuan (2011) studied properties of the L_1 penalized SVM. Wang et al. (2006) proposed SVM with a combination of L_1 and L_2 penalties. Zou and Yuan (2008) considered the L_∞ penalized SVM when there is prior knowledge about the grouping information of features. Zhang et al. (2006) and Becker et al. (2011) suggested SVM with a non-convex penalty in the application of gene selection. Though all these methods target selecting the best subset of features, theoretical justification about how well the selected subset is estimating the true model is still largely underdeveloped. Recently Zhang et al. (2014) showed that the SVM penalized with a class of non-convex penalties enjoys the oracle property (Fan and Li, 2001), that is, the estimated classifier behaves as if the subset of all relevant features is known *a priori*. Yet this model selection consistency result relies heavily on the proper choice of the involved tuning parameter which is often selected by cross-validation in practice. However, Wang et al. (2007) showed that the generalized cross-validation criterion can lead to overfitting even with a very large sample size.

Information criteria such as AIC (Akaike, 1973) and BIC (Schwarz, 1978) have been used for model selection and their theoretical properties have been well studied, see Shao (1997), Shi and Tsai (2002) and references therein. It is well understood that the BIC can identify the true model consistently when the dimensionality is fixed. The idea of combining information criterion with support vector machine to select relevant features was first proposed in Claeskens et al. (2008). They proposed the SVM information criterion (SVMIC_L) and provided some encouraging numerical evidence. Yet its theoretical properties, such as model selection consistency, have not been investigated.

In this paper, we propose a consistent SVM information criterion for model selection in the diverging model spaces. We first fill the gap by providing theoretical justification for the criterion SVMIC_L proposed in Claeskens et al. (2008). Our results show that this information criterion is model selection consistent in the fixed dimensional model space, but it can be too liberal when the candidate model space is diverging. To remedy this problem, a modified information criterion for high dimensional case (SVMIC_H) is introduced. The extension of model selection consistency from SVMIC_L to SVMIC_H is a challenging problem. The point-wise consistency of SVM solution is enough to justify the model selection consistency if the number of candidate models is fixed. Nevertheless, in the diverging model

spaces the probabilities for favoring an underfitted or overfitted model by the information criterion can accumulate at a very fast speed and alternative techniques are required. We develop the uniform consistency of SVM solution which has not been carefully studied in the literatures. Based on the uniform convergence rate, we prove that the new information criterion possesses model selection consistency even when the number of features diverges at an exponential rate of the sample size. That is, with probability arbitrarily close to one, we can identify the true model from all the underfitted or overfitted models in the diverging model spaces. To the best of our knowledge, this is the first result of model selection consistency for the SVM. We further apply this information criterion to the problem of tuning parameter selection in penalized SVMs. The proposed support vector machine information criterion can be computed easily after fitting the SVM with computation cost much lower than resampling methods like cross-validation. Simulation studies and real data examples confirm the superior performance of the proposed method in terms of model selection consistency and computational scalability.

In Section 2 we define the support vector machine information criterion. Its theoretical properties are studied in Section 3. Sections 4 and 5 present numerical results on simulation examples and real-world gene selection datasets, respectively. We conclude with some discussions in Section 6.

2. Support vector machine information criterion

In this paper we use normal font for scalars and bold font for vectors or matrices. Consider a random pair $(\mathbf{X}, Y)$ with $\mathbf{X}^T = (1, X_1, \ldots, X_p) = (1, (\mathbf{X}^+)^T) \in \mathbf{R}^{(p+1)}$ and $Y \in \{1, -1\}$. Let $\{(\mathbf{X}_i, Y_i)\}_{i=1}^n$ be a set of training data independently drawn from the distribution of $(\mathbf{X}, Y)$. Denote $\boldsymbol{\beta}$ to be a $(p+1)$-dimensional vector of interest with $\boldsymbol{\beta}^T = (\beta_0, \beta_1, \ldots, \beta_p) = (\beta_0, (\boldsymbol{\beta}^+)^T) \in \mathbf{R}^{(p+1)}$. Let $||\cdot||$ be the Euclidean norm operator of a vector. The goal of linear SVM is to estimate a hyperplane defined by $\mathbf{X}^T\boldsymbol{\beta} = 0$ via solving the optimization problem

$$\min_{\boldsymbol{\beta}} \{C \sum_{i=1}^{n} \xi_i + \frac{1}{2}||\boldsymbol{\beta}^+||^2\} \tag{1}$$

subject to the constraints that $\xi_i \geq 0$ and $Y_i\mathbf{X}_i^T\boldsymbol{\beta} \geq 1 - \xi_i$ for all $i = 1, \ldots, n$, where $C > 0$ is a tuning parameter. This can be written equivalently into an unconstrained regularized empirical loss minimization problem:

$$\min_{\boldsymbol{\beta}} \{\frac{1}{n} \sum_{i=1}^{n} H(Y_i\mathbf{X}_i^T\boldsymbol{\beta}) + \frac{\lambda_n}{2}||\boldsymbol{\beta}^+||^2\}, \tag{2}$$

where $H(t) = (1-t)_+$ is the hinge loss function, $(z)_+ = \max(z, 0)$ and $\lambda_n > 0$ is a tuning parameter with $C = (n\lambda_n)^{-1}$.

Following the definition in Koo et al. (2008), we denote $(\boldsymbol{\beta}^*)^T = (\beta_0^*, \beta_1^*, \ldots, \beta_p^*) = (\beta_0^*, (\boldsymbol{\beta}^{*+})^T) \in \mathbf{R}^{(p+1)}$ to be the true parameter value that minimizes the population hinge loss. That is,

$$\boldsymbol{\beta}^* = \arg\min_{\boldsymbol{\beta}} \mathbb{E}(1 - Y\mathbf{X}^T\boldsymbol{\beta})_+.$$

Note that $\boldsymbol{\beta}^*$ is not necessarily always the same as the Bayes rule. However, it gives the optimal upper bound of the risk of the 0-1 loss through convex relaxation and its sparsity structure is exactly the same as the one of Bayes rule in special cases such as linear discriminant analysis. For more discussions see Zhang et al. (2014). Denote $S = \{j_1, \ldots, j_d\} \subset \{1, \ldots, p\}$ to be a candidate model, $\mathbf{X}_{i,S}^T = (1, X_{i,j1}, \ldots, X_{i,jd})$, $\boldsymbol{\beta}_S^T = (\beta_0, \beta_{j1}, \ldots, \beta_{jd})$ and $|S|$ the cardinality of S. The subset of all relevant features is defined by $S^* = \{j : 1 \leq j \leq p, \beta_j^* \neq 0\}$. We assume that the truth $\boldsymbol{\beta}^*$ is sparse (i.e., most of its components are exactly zero). Denote $q = |S^*|$ which characterizes the sparsity level. We assume that q is fixed and does not depend on n. In this paper, we consider the diverging model spaces in which the dimensionality $p = p_n$ is allowed to increase with n and can be potentially much larger than n. We also assume that $\lambda_n \to 0$ as $n \to \infty$ and only consider the non-separable case in the limit to ensure the uniqueness of the truth $\boldsymbol{\beta}^*$. Here by non-separable, we mean that the two classes cannot be linearly separated from each other.

To identify the true model S^*, Claeskens et al. (2008) proposed an information criterion for SVM (denoted by SVMIC$_L$) based on the slack variables $\{\xi_i\}_{i=1}^n$. That is,

$$\text{SVMIC}_L(S) = \sum_{i=1}^n \xi_i + |S| \log(n),$$

where $\{\xi_i\}_{i=1}^n$ are obtained from (1) only using the variables in S. This information criterion is equivalent to

$$\text{SVMIC}_L(S) = \sum_{i=1}^n (1 - Y_i \mathbf{X}_{i,S}^T \widehat{\boldsymbol{\beta}}_S)_+ + |S| \log(n), \tag{3}$$

where $\widehat{\boldsymbol{\beta}}_S = \arg\min\{1/n \sum_{i=1}^n (1 - Y_i \mathbf{X}_{i,S}^T \boldsymbol{\beta}_S)_+ + \lambda_n/2 ||\boldsymbol{\beta}_S^+||^2\}$. It is evident that the SVMIC$_L$ directly follows the spirit of BIC. Claeskens et al. (2008) fixed $C = 1$ in (1) and found minor difference for different choices of C, which is equivalent to $\lambda_n = 1/n$ in (2). To be consistent with the work in Claeskens et al. (2008), we also consider this choice of λ_n in this paper. There are two potential drawbacks of this information criterion. First, though supported with numerical findings, theoretical properties of SVMIC$_L$, such as model selection consistency, are largely unknown even under the assumption of a fixed p. Second, in many real world datasets where the dimension can be much larger than the sample size, it would be more appropriate to consider the model selection problem in the framework of diverging model spaces. This extension from low dimensions to high dimensions can greatly change the theoretical properties of the information criterion. Chen and Chen (2008) showed that the ordinary BIC for linear regression cannot identify the true model consistently in the diverging p case. Wang et al. (2009) showed that the ordinary BIC fails to select a consistent shrinkage level in penalized least squares regression with a diverging p. Such results in the literature suggest that SVMIC$_L$ may also suffer from inconsistency in high dimensions and alternative criterion is needed.

To overcome these issues, we propose a modified support vector machine information criterion for model selection in a high dimensional model space (denote by SVMIC$_H$). This criterion is adpated from SVMIC$_L$ and defined as

$$\text{SVMIC}_H(S) = \sum_{i=1}^n (1 - Y_i \mathbf{X}_{i,S}^T \widehat{\boldsymbol{\beta}}_S)_+ + L_n |S| \log(n), \tag{4}$$

where $\widehat{\boldsymbol{\beta}}_S = \arg\min\{1/n\sum_{i=1}^n(1-Y_i\mathbf{X}_{i,S}^T\boldsymbol{\beta}_S)_+ + \lambda_n/2||\boldsymbol{\beta}_S^+||^2\}$ and L_n is a constant sequence that diverges to infinity. Note that if L_n is a non-diverging constant then this reduces to SVMIC_L in the limit. We will show that SVMIC_H possesses the nice property of model selection consistency even when p increases at an exponential rate of n. Compared to SVMIC_L in Claeskens et al. (2008), our information criterion SVMIC_H adds larger penalty to the size of the selected subset and behaves more conservatively. As we will see, this additional preference for simpler models plays an important role in consistent model selection when we are searching over diverging model spaces.

We make two remarks about SVMIC_H. First, the choice of L_n in (4) is flexible. It is *not* a tuning parameter and does not need to be chosen by computationally intensive methods such as cross-validation. We will show that a wide spectrum of L_n can lead to a consistent information criterion. This is further confirmed in our simulations and real data analysis where we examine different choices of L_n. Therefore the computation cost of SVMIC_H is the same as SVMIC_L and much lower than cross-validation. Second, it is possible to define the information criterion as the log-transformed version, that is

$$\log(\sum_{i=1}^n(1-Y_i\mathbf{X}_{i,S}^T\widehat{\boldsymbol{\beta}}_S)_+) + L_n|S|\log(n)/n$$

which is scalar invariant. It can be shown the model selection consistency still holds for this definition. However, we follow the advice from Guyon et al. (2002) to standardize variables before training SVM and as a consequence we automatically have scalar invariance of the sum of slack variables. To be consistent with SVMIC_L defined in Claeskens et al. (2008), we take definition (4) in our paper.

3. Theoretical results

3.1 Notations and conditions

To facilitate technical proofs, we introduce some additional notation. Denote $L(\boldsymbol{\beta}) = \mathbb{E}(1-Y\mathbf{X}^T\boldsymbol{\beta})_+$. Recall that $\boldsymbol{\beta}^* = \arg\min_{\boldsymbol{\beta}} L(\boldsymbol{\beta})$. Let $\mathbf{S}(\boldsymbol{\beta}) = -\mathbb{E}[\mathbf{1}(1-Y\mathbf{X}^T\boldsymbol{\beta}\geq 0)Y\mathbf{X}]$, where $\mathbf{1}(\cdot)$ is the indicator function. Also define $\mathbf{H}(\boldsymbol{\beta}) = \mathbb{E}[\delta(1-Y\mathbf{X}^T\boldsymbol{\beta})\mathbf{X}\mathbf{X}^T]$, where $\delta(\cdot)$ is the Dirac delta function. Koo et al. (2008) showed that under some regularity conditions, $\mathbf{S}(\boldsymbol{\beta})$ and $\mathbf{H}(\boldsymbol{\beta})$ behave like the gradient and Hessian matrix of $L(\boldsymbol{\beta})$, respectively. Furthermore we denote f_+ and f_- to be the densities of $\mathbf{X}^+ \in \mathbf{R}^p$ conditioning on $Y=1$ and $Y=-1$, respectively.

Given the dimension p, the number of candidate models is 2^p-1. When p is very large, we cannot afford to calculate $\text{SVMIC}_H(S)$ for all possible subsets. Instead, we only search for the best model in a restricted model space. To be more specific, we denote $\widehat{S}$ the model chosen by SVMIC_H such that

$$\widehat{S} = \arg\min_{S:|S|\leq M_n}\text{SVMIC}_H(S), \tag{5}$$

where M_n is a sequence of positive integers that bounds the size of the restricted model space from above. In this paper, we consider $M_n = O(n^\kappa)$ for some constant $0<\kappa<1/2$, that is, we only consider the candidate model with the size diverges slower than $\sqrt{n}$. One motivation for this choice of M_n is the "bet on sparsity" principle (Hastie et al., 2001).

Note that by Lemma 1 of Zhang et al. (2014), if the number of relevant features diverges faster than $\sqrt{n}$, the true parameter $\boldsymbol{\beta}^*$ cannot be estimated consistently even with the oracle information of the true model S^*. Therefore, there is no need to consider those models with sizes increasing faster than $\sqrt{n}$ as in general no method would work for them even when the true underlying model is known. Notice also that it is possible to prove the model selection consistency without the restricted model space. However, this would requires $p_n = o(\sqrt{n})$, which cannot diverge very fast with the sample size.

We now present the technical conditions that are needed for studying the theoretical properties of SVMIC$_H$.

(A1) f_+ and f_- are continuous and have some common support in $\mathbf{R}^p$.

(A2) $|X_j| \leq M < \infty$ for some positive constant M and $1 \leq j \leq p$.

(A3) For all $S \in \{S : |S| \leq M_n, S \supseteq S^*\}$, $\lambda_{\max}(\mathbb{E}(\mathbf{X}_{i,S}\mathbf{X}_{i,S}^T)) \leq c_1$, where $\lambda_{\max}(\cdot)$ is the largest eigenvalue of a matrix and $c_1 > 0$ is a constant.

(A4) The densities of $\mathbf{X}_{i,S^*}^T\boldsymbol{\beta}_{S^*}^*$ conditioning on $Y = 1$ and $Y = -1$ are uniformly bounded away from zero and infinity at the neighborhood of $\mathbf{X}_{i,S^*}^T\boldsymbol{\beta}_{S^*}^* = 1$ and $\mathbf{X}_{i,S^*}^T\boldsymbol{\beta}_{S^*}^* = -1$, respectively.

(A5) $M_n = O(n^\kappa)$ for some constant $0 < \kappa < 1/2$.

(A6) $p_n = O(\exp(n^\gamma))$ for some constant $0 < \gamma < (1 - 2\kappa)/5$.

(A7) For all $S \in \{S : |S| \leq M_n, S \supseteq S^*\}$, there exist some positive constants c_2 and c_3 such that $\lambda_{\min}(\mathbf{H}(\boldsymbol{\beta}_S)) \geq c_2$ and $\lambda_{\max}(\mathbf{H}(\boldsymbol{\beta}_S)) = O(|S|)$ over the set $\{\boldsymbol{\beta} : ||\boldsymbol{\beta} - \boldsymbol{\beta}^*|| \leq c_3\}$, where $\lambda_{\min}(\cdot)$ is the smallest eigenvalue of a matrix.

(A8) For all $S \in \{S : |S| \leq M_n, S \supseteq S^*\}$, $\lambda_{\max}(\mathbf{H}(\boldsymbol{\beta}_S^*)) \log(p_n) = o(L_n \log(n))$, $L_n \log(n) = o(n)$.

Conditions (A1) is required so that $\mathbf{S}(\boldsymbol{\beta})$ and $\mathbf{H}(\boldsymbol{\beta})$ are well-defined, see Koo et al. (2008) for more details. Condition (A2) is assumed in the literature of high dimensional model selection consistency as in Wang et al. (2012) and Lee et al. (2014). Condition (A3) on the largest eigenvalue is similar to the sparse Riesz condition (Zhang and Huang, 2008) and is often assumed for model selection consistency in the diverging p scenario (Chen and Chen, 2008; Yuan, 2010; Zhang, 2010). Note that the lower bound on the eigenvalue of the covariance matrix of $\mathbf{X}_S$ is not specified. Condition (A4) assumes that as the sample size increases, there is enough information around the non-differentiable point of the hinge loss function. This condition is also required for model selection consistency of non-convex penalized SVM in high dimensions (Zhang et al., 2014). Condition (A6) specifies that p is allowed to diverge at an exponential rate of n. Conditions (A7) requires that the Hessian matrix is well-behaved. More specifically, (A7) requires a lower bound on the smallest eigenvalue of the Hessian matrix in the neighborhood of the true value. Koo et al. (2008) gave sufficient conditions for the positive-definiteness of the Hessian matrix at the true value and showed these conditions hold under the setting of Fisher's linear discriminant analysis with a fixed model size. Condition (A8) specifies the rate requirement of L_n in (4).

3.2 Consistency of SVMIC$_L$ for a fixed p

In this section we assume that the dimension p is fixed and study the theoretical properties of SVMIC$_L$. Let $\Omega = \{S : |S| \leq M\}$ be the candidate model space where M is a positive number. Furthermore, when p is fixed, the total number of candidate models is also fixed.

Note that, to prove the model selection consistency of SVMIC_L, we need to show that

$$\Pr(\inf_{S\in\Omega, S\neq S^*} \text{SVMIC}_L(S) > \text{SVMIC}_L(S^*)) \to 1$$

as $n \to \infty$. By the fact that Ω is a fixed model space, it is sufficient to show the result point-wisely in the model space, that is,

$$\Pr(\text{SVMIC}_L(S) > \text{SVMIC}_L(S^*)) \to 1 \tag{6}$$

for every $S \in \{S : S \in \Omega, S \neq S^*\}$. This point-wise version greatly simplifies the proof. Recall that $\widehat{\boldsymbol{\beta}}_S = \arg\min\{1/n \sum_{i=1}^n (1 - Y_i \mathbf{X}_{i,S}^T \boldsymbol{\beta}_S)_+ + \lambda_n/2 ||\boldsymbol{\beta}_S^+||^2\}$. As we will show in the appendix, it suffices to conclude (6) with the condition that $\widehat{\boldsymbol{\beta}}_S$ is a consistent estimator of $\boldsymbol{\beta}_S^*$ whenever the candidate subset S includes the true subset S^*. By Theorem 1 of Koo et al. (2008), we have $||\widehat{\boldsymbol{\beta}}_S - \boldsymbol{\beta}_S^*|| = O_p(n^{-1/2})$ for every fixed $S \supset S^*$ under the assumption of fixed p. Therefore the point-wise consistency holds. The result is summarized in Lemma 1.

Lemma 1 *Assuming p is a fixed number and $\lambda_n = 1/n$. Under conditions (A1)-(A4) and (A6)-(A7), we have*

$$\Pr(\widehat{S} = S^*) \to 1$$

as $n \to \infty$, where $\widehat{S} = \arg\min_{S:|S|\leq M} SVMIC_L(S)$.

3.3 Consistency of SVMIC_H for a diverging p

The proof becomes much more involved when p is diverging, especially when p diverges much faster than $O(\sqrt{n})$. Let $\Omega = \{S : |S| \leq M_n\}$, $\Omega_+ = \{S : |S| \leq M_n, S \supset S^*, S \neq S^*\}$ and $\Omega_- = \{S : |S| \leq M_n, S \not\supset S^*\}$, where Ω_+ and Ω_- are spaces of overfitted and underfitted models, respectively. Though the information criterion for the fixed model space can differentiate the true model from an arbitrary candidate model, this point-wise result is not sufficient for the overall consistency if the problem requires searching uniformly over a diverging model space. That is, even if (6) holds for every $S \in \Omega$, we still cannot conclude model selection consistency, as the probability of favoring an overfitted or underfitted candidate model rather than the true model can accumulate at very fast speed if the number of candidate models is diverging and hence lead to inconsistent model selection. To control the overall failing probability, we need a uniform convergence rate of SVM solution $\widehat{\boldsymbol{\beta}}_S$ over the diverging model space Ω. Note that $\Omega = \Omega_+ \cup \{S^*\} \cup \Omega_-$. It turns out that the uniform convergence rate of $\widehat{\boldsymbol{\beta}}_S$ over $S \in \Omega_+$ is sufficient for the technical proof. We summarize the uniform rate in Lemma 2 below.

Lemma 2 *Under conditions (A1)-(A7) and $\lambda_n = 1/n$, we have*

$$\sup_{S:|S|<M_n, S\supset S^*} ||\widehat{\boldsymbol{\beta}}_S - \boldsymbol{\beta}_S^*|| = O_p(\sqrt{|S|\log(p)/n}).$$

This uniform convergence rate of SVM solution is far from being a trivial result. Recently Zhang et al. (2014) showed that $||\widehat{\boldsymbol{\beta}}_S - \boldsymbol{\beta}_S^*|| = O_p(\sqrt{|S|/n})$ for a specific diverging model S which satisfies $S \supset S^*$ and $|S| = o(\sqrt{n})$. Although it is an extension of Theorem 1 in Koo

et al. (2008) to the diverging p case, it is still only a point-wise result and cannot be applied directly to bound the overall failing probability. Not surprisingly, the uniform convergence rate in Lemma 2 is slower by a factor $\sqrt{\log(p)}$, which is the price we pay to search over the candidate model space uniformly. In fact, this additional term is the main reason for adding the extra penalty L_n in SVMIC_H.

We now give an intuitive explanation why SVMIC_L can fail in the diverging model space. Consider all the overfitted models in Ω_+. We have the following decomposition

$$\begin{aligned}&\inf_{S\in\Omega_+} \text{SVMIC}_L(S) - \text{SVMIC}_L(S^*)\\ =&\inf_{S\in\Omega_+} \big\{\sum_{i=1}^n (1 - Y_i \mathbf{X}_{i,S}^T \widehat{\boldsymbol{\beta}}_S)_+ - \sum_{i=1}^n (1 - Y_i \mathbf{X}_{i,S^*}^T \boldsymbol{\beta}_S^*)_+ + (|S| - |S^*|)\log(n)\big\}. \end{aligned} \tag{7}$$

Note that the difference of the sum of hinge loss can be negative and the difference of model size is always positive. We will show in the appendix that the difference of the sum of hinge loss is of order $O_p(n||\widehat{\boldsymbol{\beta}}_S - \boldsymbol{\beta}_S^*||^2)$ under some regularity conditions. For fixed p, it implies that the difference of model size dominates for large n and the sign of (7) is always positive in the limit. For diverging p, however, this is not always the case. By Lemma 2, the difference of hinge loss in (7) is of order $O(|S|\log(p))$ and the difference of model size is of order $O(|S|\log(n))$, thus the sign of (7) can be negative in the limit. Therefore even if we have a very large sample size, SVMIC_L may still favoring the models that are overfitted due to the slower uniform convergence rate of SVM solution. Because SVMIC_L can be viewed as directly following the spirit of ordinary BIC, this result agrees with the findings reported in Chen and Chen (2008) that BIC can be too liberal in high dimensional model selection. Here by liberal, we mean that there is a positive probability that an overfitted model is more favored than the true model by the information criterion even with an infinite sample size.

For SVMIC_H, we can do a similar decomposition as in (7) for all the overfitted models

$$\begin{aligned}&\inf_{S\in\Omega_+} \text{SVMIC}_H(S) - \text{SVMIC}_H(S^*)\\ =&\inf_{S\in\Omega_+} \big\{\sum_{i=1}^n (1 - Y_i \mathbf{X}_{i,S}^T \widehat{\boldsymbol{\beta}}_S)_+ - \sum_{i=1}^n (1 - Y_i \mathbf{X}_{i,S^*}^T \boldsymbol{\beta}_S^*)_+ + L_n(|S| - |S^*|)\log(n)\big\}. \end{aligned} \tag{8}$$

Note that the extra term L_n diverges to infinity and the sign of (8) in the limit is determined by the difference of model size, which is always positive. That is, SVMIC_H can identify the true model from all the overfitted models for sufficiently large n. This result is summarized in Lemma 3.

Lemma 3 *Under conditions (A1)-(A8) and $\lambda_n = 1/n$, we have*

$$\Pr(\inf_{S:S\in\Omega_+} SVMIC_H(S) > SVMIC_H(S^*)) \to 1.$$

as $n \to \infty$.

To conclude the model selection consistency, we also need to consider all the underfitted models. This requires a different analysis because for every underfitted model $S \in \Omega_-$, the

difference of the model size $|S| - |S^*|$ can be negative and thus the decomposition in (7) is not helpful. However, one can always add relevant features to the underfitted model and study the enlarged model instead. To be more specific, for every $S \in \Omega_-$, one can always create the enlarged model $\tilde{S}$ such that $\tilde{S} = S \cup S^*$. Note that $\tilde{S}$ is either an overfitted model or the true model. The model $\tilde{S}$ which includes all the signals bridges the underfitted and overfitted model space through the simple fact

$$\begin{aligned}&\inf_{S:S\in\Omega_-} \mathrm{SVMIC}_H(S) - \mathrm{SVMIC}_H(S^*)\\ =&\inf_{S:S\in\Omega_-} \{[\mathrm{SVMIC}_H(S) - \mathrm{SVMIC}_H(\tilde{S})] + [\mathrm{SVMIC}_H(\tilde{S}) - \mathrm{SVMIC}_H(S^*)]\}.\end{aligned}$$

By Lemma 3, in the limit the difference $\mathrm{SVMIC}_H(\tilde{S}) - \mathrm{SVMIC}_H(S^*)$ is non-negative for every $S \in \Omega_-$ (could be exactly 0). Note also that the difference between S and $\tilde{S}$ is at least one missing relevant feature. According to the assumption that the signals do not diminish to 0 as sample size increases, one can show the model with more signals always produces a strictly smaller sum of hinge loss in the limit. That is, for sufficiently large n, we always have

$$\sum_{i=1}^{n}(1 - Y_i\mathbf{X}_{i,S}^T\widehat{\boldsymbol{\beta}}_S)_+ - \sum_{i=1}^{n}(1 - Y_i\mathbf{X}_{i,\tilde{S}}^T\widehat{\boldsymbol{\beta}}_{\tilde{S}})_+ \geq C > 0$$

for every $S \in \Omega_-$ and some constant C does not depend on S. Then we arrive at the following result for the underfitted model space.

Lemma 4 *Under conditions (A1)-(A8) and $\lambda_n = 1/n$, we have*

$$\Pr(\inf_{S:S\in\Omega_-} SVMIC_H(S) > SVMIC_H(S^*)) \to 1.$$

as $n \to \infty$.

By combing Lemma 3 and Lemma 4, we can conclude the model selection consistency of SVMIC_H in the diverging model space.

Theorem 5 *Under conditions (A1)-(A8) and $\lambda_n = 1/n$, we have*

$$\Pr(\hat{S} = S^*) \to 1.$$

as n, $p \to \infty$, where $\widehat{S} = \arg\min_{S:|S|\leq M_n} SVMIC_H(S)$.

3.4 Application to tuning parameter selection in penalized SVMs

Theorem 1 states that SVMIC_H can identify the true model from Ω. However, in practice it can be very time-consuming and even infeasible to calculate $\mathrm{SVMIC}_H(S)$ for every $S \in \Omega$. One possible approach is to form a solution path via penalized SVM and only consider the candidate models on the path. The idea of using solution path has been shown to greatly reduce the computation burden, see Mazumder et al. (2011). For the solution path of penalized SVM, Hastie et al. (2004) studied the L_1 penalized SVM and showed that the solution path is piece-wise linear in C which is the regularization parameter in (1).

Model selection on the solution path is essentially a tuning parameter selection problem. Recently, several methods have been proposed for choosing the tuning parameter based

on the BIC-type information criterion, including Wang et al. (2009) for penalized linear regression, Kawano (2012) for bridge regression, Lee et al. (2014) for penalized quantile regression and Fan and Tang (2013) for penalized generalized linear model. Following the ideas therein, we propose to choose the shrinkage level of penalized SVMs based on the modified support vector machine information criterion. Let $\widehat{\boldsymbol{\beta}}_{\lambda_n}^T = (\hat{\beta}_{\lambda_n,0}, \hat{\beta}_{\lambda_n,1}, \ldots, \hat{\beta}_{\lambda_n,p})$ be the solution to some penalized SVM with a tuning parameter λ_n. That is,

$$\widehat{\boldsymbol{\beta}}_{\lambda_n} = \arg\min_{\boldsymbol{\beta}} \Big\{ \frac{1}{n} \sum_{i=1}^{n} (1 - Y_i \mathbf{X}_i^T \boldsymbol{\beta})_+ + \sum_{j=1}^{p} p_{\lambda_n}(\beta_j) \Big\}, \tag{9}$$

where $p_{\lambda_n}(\cdot)$ is some penalty function with a tuning parameter λ_n. Denote $\widehat{S}_{\lambda_n} = \{j : 1 \leq j \leq p, \hat{\beta}_{\lambda_n,j} \neq 0\}$. We define the information criterion for choosing tuning parameter λ_n as

$$\text{SVMIC}_H(\lambda_n) = \sum_{i=1}^{n} (1 - Y_i \mathbf{X}_i^T \widehat{\boldsymbol{\beta}}_{\lambda_n})_+ + L_n |\widehat{S}_{\lambda_n}| \log(n),$$

where L_n is defined in $\text{SVMIC}_H(S)$. The selected tuning parameter is the one that minimizes the information criterion and results in the model size within the restricted model space, that is,

$$\widehat{\lambda}_n = \underset{\lambda : |\widehat{S}_\lambda| \leq M_n}{\arg\min} \text{SVMIC}_H(\lambda).$$

This information criterion for selecting tuning parameter can be applied to various penalized approaches for sparse SVMs. Note that the feature selection consistency of the SCAD penalized SVM is shown to rely on the proper choice of the tuning parameter (Zhang et al., 2014), where resampling procedure such as five-fold cross-validation is commonly used in practice. As we will also show in our numerical findings in Section 4.2, the proposed information criterion $\text{SVMIC}_H(\lambda_n)$ usually select the shrinkage level that leads to the correct model size. The tuning parameter selected by cross-validation, however, is more likely to be under-penalized and lead to an overfitted model. Furthermore, the cross-validation is more computationally intensive than information criterion and hence less desirable when the number of features is large.

4. Simulations

In this section we study the finite-sample performance of SVMIC_H. We are interested in the model selection ability of $\text{SVMIC}_H(S)$ and the tuning parameter selection ability of $\text{SVMIC}_H(\lambda_n)$. We also examine the effect of different choices of L_n in the definition of SVMIC_H. For all simulations, we consider the rates $\log(\log(n))$, $\sqrt{\log(n)}$, $\log(n)$ and $n^{-1/3}$ for L_n. We compare with SVMIC_L in Claeskens et al. (2008) and the extended Bayesian information criterion (EBIC) proposed in Chen and Chen (2008). Note that EBIC is originally proposed for model selection in the diverging model space in the framework of regression and it has not been applied into classification problem. However, the main strategy therein is to add an additional term $\log(\binom{p}{|S|}) \log(n)$ in the BIC penalty, where $\binom{p}{|S|}$ is the number of $|S|$ combinations chosen from p items. We modify EBIC for model

selection of SVM by selecting the model S that minimizes the criterion

$$\sum_{i=1}^{n}(1-Y_i\mathbf{X}_{i,S}^T\widehat{\boldsymbol{\beta}}_S)_+ + |S|\log(n) + \log\left(\binom{p}{|S|}\right)\log(n).$$

This modification essentially follows the idea in Chen and Chen (2008) and we are interested in its finite-sample performance compared with SVMIC$_H$.

To investigate these issues, we conduct the SCAD penalized SVM, which has been shown to enjoy the model selection consistency for a properly chosen tuning parameter (Zhang et al., 2014). That is, given a specific λ_n, we solve (9) with $p_{\lambda_n}(\cdot)$ being the SCAD penalty defined in Fan and Li (2001). The corresponding optimization problem is a non-convex one, for which the local linear approximation (LLA) algorithm (Zou and Li, 2008) is implemented in all our numerical studies. To be more specific, for step $t \geq 1$, given the solution $\widehat{\boldsymbol{\beta}}^{(t-1)} = (\hat{\beta}_0^{(t-1)}, \ldots, \hat{\beta}_p^{(t-1)})^T$ at the previous step, we update by solving

$$\widehat{\boldsymbol{\beta}}^{(t)} = \arg\min_{\boldsymbol{\beta}}\{\frac{1}{n}\sum_{i=1}^{n}(1-Y_i\mathbf{X}_i^T\boldsymbol{\beta})_+ + \sum_{j=1}^{p}p'_{\lambda_n}(|\hat{\beta}_j^{(t-1)}|)|\beta_j|\},$$

where $p'_{\lambda_n}(\cdot)$ denotes the derivative of $p_{\lambda_n}(\cdot)$. Note that each update step can be easily recast as a linear programming (LP) problem and efficiently solved by many popular solvers. In this paper we take the initial value $\{\widehat{\boldsymbol{\beta}}^{(0)} : \hat{\beta}_j^{(0)} = 0, 0 \leq j \leq p\}$ and claim convergence if the value $||\widehat{\boldsymbol{\beta}}^{(t-1)} - \widehat{\boldsymbol{\beta}}^{(t)}||$ is small enough.

4.1 Model selection of SVMIC$_H(S)$

In this subsection we study the model selection ability of SVMIC$_H(S)$. The data are generated from two models. The first model is adapted from Fisher's linear discriminant analysis (LDA) and the second model is related to probit regression.

- Model 1: $\Pr(Y=1) = \Pr(Y=-1) = 0.5$, $\mathbf{X}|(Y=1) \sim MN(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, $\mathbf{X}|(Y=-1) \sim MN(-\boldsymbol{\mu}, \boldsymbol{\Sigma})$, $q=4$, $\boldsymbol{\mu} = (0.25, 0.25, 0.25, 0.25, 0, \ldots, 0)^T \in \mathbf{R}^p$, $\boldsymbol{\Sigma} = (\sigma_{ij})$ with nonzero elements $\sigma_{ii} = 1$ for $i = 1, 2, \cdots, p$ and $\sigma_{ij} = \rho = -0.2$ for $1 \leq i \neq j \leq q$. The Bayes rule is given by sign$(X_1+X_2+X_3+X_4)$ with Bayes error 21.4%.

- Model 2: $\mathbf{X} \sim MN(\mathbf{0}_p, \boldsymbol{\Sigma})$, $\boldsymbol{\Sigma} = (\sigma_{ij})$ with nonzero elements $\sigma_{ii} = 1$ for $i = 1, 2, \cdots, p$ and $\sigma_{ij} = \rho = 0.4^{|i-j|}$ for $1 \leq i \neq j \leq q$, $\Pr(Y=1) = \Phi(\mathbf{X}^T\boldsymbol{\beta})$ where $\Phi(\cdot)$ is the CDF of the standard normal distribution, $q=4$, $\boldsymbol{\beta} = (0.8, 0.8, 0.8, 0.8, 0, \ldots, 0)^T$. The Bayes rule is sign$(0.57X_1+0.34X_2+0.34X_3+0.57X_4)$ with Bayes error 11.5%.

For both models, we construct the solution path using SCAD penalized SVM for candidate models with $|S| \leq M_n = 50$. Our goal is to check how different information criteria evaluate and select the optimal model from all the candidate models on the solution path. We consider three different (n, p) combinations with p ranging from 2000 to 4000 and n is only one tenth of p. Note that Model 1 is a very noisy model with high Bayes error and Model 2 is less noisy but with moderate correlation between the relevant features. We use 200 replications to see the variations of the results. The columns "Correct", "Underfit" and

"Overfit" summarize the percentages over 200 replications for correct model selection, overfitting and underfitting, respectively. The numbers under columns "Signal" and "Noise" are the average numbers of selected relevant and irrelevant features, respectively. We also generate an independent dataset with sample size 10000 to evaluate the test error. Numbers in parentheses are the corresponding standard errors.

Table 1 summarizes the model selection results of SVMIC$_L$, SVMIC$_H$ and the criterion proposed in Chen and Chen (2008) for Model 1. For all (n, p) combinations, SVMIC$_H$ shows uniformly higher percentages to identify the correct model than SVMIC$_L$ regardless of the choices of L_n. It can be seen that in the cases p is much larger than n, SVMIC$_L$ behaves too liberal and tends to select an overfitted model. Note that SVMIC$_H$ also has a significantly lower testing error than SVMIC$_L$ in all settings even when SVMIC$_L$ includes slightly more signals in the model. This agrees with the findings in Fan and Fan (2008) that the accumulation of the noises can greatly blur the prediction power. Though the SVMIC$_H$ with different L_n all performs better than SVMIC$_L$, their performances are not exactly the same. For the criteria with a more aggressive penalty on the model size such as $\log(n)$ and $n^{-1/3}$, there are considerable underfitting when the sample size is small ($n = 200$). As the sample size increases, the difference of L_n decreases. This suggests that although asymptotically the choice of L_n can lie in a wide range of spectrums, for small sample sizes some choices of L_n can be too conservative and may not be much better than SVMIC$_L$ which is too liberal. In general, we find $L_n = \sqrt{\log(n)}$ seems to be a reasonable choice for many scenarios.

Another interesting finding is the comparison to the criterion following the spirit in Chen and Chen (2008). Though its theoretical property has not been investigated, the empirical results suggest that it performs similar to SVMIC$_H$ with $L_n = \sqrt{\log(n)}$ in finite samples. In fact, by using the approximation that $\binom{p}{|S|} \approx p^{|S|}$ when p is much larger than $|S|$, one can easily show that

$$\log(n)|S|\log(\log(n)) < \log(n)|S| + \log(n)\log\left(\binom{p}{|S|}\right) < \log(n)|S|\log(n)$$

for $n < p < 10^3 n$ and a very wide range of n. This provides some evidence that the criterion directly adapted from Chen and Chen (2008) behaves more libearal than SVMIC$_H$ with $L_n = \log(n)$ but is more aggressive than SVMIC$_H$ with $L_n = \log(\log(n))$.

Table 2 summarizes the model selection results for Model 2. The SVMIC$_H$ with $\log(\log(n))$ and $\sqrt{\log(n)}$ as L_n perform uniformly better than SVMIC$_L$ and the criterion in Chen and Chen (2008) for all scenarios. Due to the correlations among the signals, the more aggressive choices of L_n suffer from considerable underfitting. Again our empirical results suggest that $L_n = \sqrt{\log(n)}$ seems to be an appropriate choice for a wide range of problems.

4.2 Tuning parameter selection of SVMIC$_H(\lambda)$

In this subsection we examine the tuning parameter selection ability of SVMIC$_H(\lambda)$. The data is generated from the following model.

- $\Pr(Y = 1) = \Pr(Y = -1) = 0.5$, $\mathbf{X}|(Y = +1) \sim MN(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, $X|(Y = -1) \sim MN(-\boldsymbol{\mu}, \boldsymbol{\Sigma})$, $q = 5$, $\boldsymbol{\mu} = (0.1, 0.2, 0.3, 0.4, 0.5, 0 \ldots, 0) \in \mathbf{R}^p$, $\boldsymbol{\Sigma} = (\sigma_{ij}) = \mathbf{1}_p$ nonzero

Table 1: Simulation results for Model 1 over 200 replications

Method	C(%)	O(%)	U(%)	Signal	Noise	Test Error(%)
			$n = 200,\ p = 2000$			
SVMIC_L	0.0	100.0	0.0	4.0	11.1	30.7(0.3)
$\text{SVMIC}_H(L_n = \log(\log(n)))$	34.0	64.0	2.0	3.8	1.6	25.4(0.3)
$\text{SVMIC}_H(L_n = \sqrt{\log(n)})$	48.0	38.0	14.0	3.6	1.0	26.1(0.4)
$\text{SVMIC}_H(L_n = \log(n))$	31.5	23.5	45.0	3.0	0.8	29.6(0.5)
$\text{SVMIC}_H(L_n = n^{-1/3})$	29.5	23.5	47.0	2.9	0.8	29.7(0.5)
Chen&Chen	47.0	26.5	26.5	3.3	0.8	27.5(0.5)
			$n = 300,\ p = 3000$			
SVMIC_L	2.0	98.0	0.0	4.0	14.1	28.7(0.3)
$\text{SVMIC}_H(L_n = \log(\log(n)))$	69.0	31.0	0.0	4.0	0.4	22.5(0.1)
$\text{SVMIC}_H(L_n = \sqrt{\log(n)})$	93.0	6.5	0.5	4.0	0.1	22.1(0.1)
$\text{SVMIC}_H(L_n = \log(n))$	69.5	3.5	27.0	3.5	0.1	25.1(0.4)
$\text{SVMIC}_H(L_n = n^{-1/3})$	53.5	3.5	43.0	3.3	0.1	26.6(0.4)
Chen&Chen	95.0	4.5	0.5	4.0	0.1	22.1(0.1)
			$n = 400,\ p = 4000$			
SVMIC_L	4.0	96.0	0.0	4.0	12.6	26.9(0.3)
$\text{SVMIC}_H(L_n = \log(\log(n)))$	77.5	22.5	0.0	4.0	0.3	22.2(0.1)
$\text{SVMIC}_H(L_n = \sqrt{\log(n)})$	95.5	4.5	0.0	4.0	0.1	21.9(0.1)
$\text{SVMIC}_H(L_n = \log(n))$	95.0	1.0	4.0	3.9	<0.1	22.4(0.1)
$\text{SVMIC}_H(L_n = n^{-1/3})$	71.0	1.0	28.0	3.5	<0.1	25.0(0.4)
Chen&Chen	98.5	1.0	0.5	4.0	<0.1	22.0(0.1)

elements $\sigma_{ii} = 1$ for $i = 1, 2, \cdots, p$ and $\sigma_{ij} = \rho = -0.2$ for $1 \le i \ne j \le q$. The Bayes rule is $\text{sign}(2.67X_1+2.83X_2+3X_3+3.17X_4+3.33X_5)$ with Bayes error: 6.3%.

We consider $p = 2000$ and 3000 and $n = 10^{-1}p$. Once the data is generated, we construct the solution path of SCAD penalized SVM on a fine grid of λ for candidate models with $|S| \le M_n = 50$. We then choose the best λ based on the definition of $\text{SVMIC}_H(\lambda)$. Similarly as the simulations for model selection, we compare with $\text{SVMIC}_L(\lambda)$ and the criterion in Chen and Chen (2008). We also implement five-fold cross-validation (denoted by 5-CV) and an adjusted version of five-fold cross-validation version (denoted by 5-CV Adj.). The adjusted 5-CV selects the most parsimonious model with MSE less than one standard error above the regular 5-CV. It is known that the adjusted 5-CV performs better than regular 5-CV in terms of selection consistency. An independent dataset with sample size 10000 is generated to evaluate the test error. This procedure is repeated for 100 replications to study the variations of the results.

Table 3 summarizes the tuning parameter selection results. It can be seen that the tuning parameter selected by SVMIC_L often leads to seriously overfitted models. As the sample size increases, SVMIC_H with all choices of L_n have a much higher chance to identify the correct model than SVMIC_L. The tuning parameter selected by SVMIC_H with $L_n = \sqrt{\log(n)}$ seems to give the most appropriate level of regularization to the model. It is not surprising that this SVMIC_H leads to great prediction power in these high dimensional cases. The

Table 2: Simulation results for Model 2 over 200 replications

Method	C(%)	O(%)	U(%)	Signal	Noise	Test Error(%)
		$n = 200,\ p = 2000$				
SVMIC_L	3.0	96.0	1.0	3.8	5.2	19.2(0.2)
$\text{SVMIC}_H(L_n = \log(\log(n)))$	31.0	31.5	37.5	3.3	0.5	17.2(0.2)
$\text{SVMIC}_H(L_n = \sqrt{\log(n)})$	25.0	2.5	72.5	3.0	0.1	18.0(0.2)
$\text{SVMIC}_H(L_n = \log(n))$	0.0	0.0	100.0	1.9	0.0	22.1(0.2)
$\text{SVMIC}_H(L_n = n^{-1/3})$	0.0	0.0	100.0	1.9	0.0	22.1(0.2)
Chen&Chen	1.0	0.0	99.0	2.2	0.0	20.4(0.2)
		$n = 300,\ p = 3000$				
SVMIC_L	6.5	93.5	0.0	4.0	7.4	18.0(0.2)
$\text{SVMIC}_H(L_n = \log(\log(n)))$	65.0	23.0	12.0	3.8	0.3	15.0(0.1)
$\text{SVMIC}_H(L_n = \sqrt{\log(n)})$	70.5	4.0	25.5	3.7	<0.1	15.3(0.1)
$\text{SVMIC}_H(L_n = \log(n))$	0.0	0.0	100.0	2.0	0.0	21.0(0.1)
$\text{SVMIC}_H(L_n = n^{-1/3})$	0.0	0.0	100.0	2.0	0.0	21.5(0.2)
Chen&Chen	33.5	0.5	66.0	3.1	<0.1	17.3(0.2)
		$n = 400,\ p = 4000$				
SVMIC_L	9.0	91.0	0.0	4.0	6.9	17.2(0.2)
$\text{SVMIC}_H(L_n = \log(\log n))$	82.0	16.5	1.5	4.0	0.2	14.5(0.1)
$\text{SVMIC}_H(L_n = \sqrt{\log(n)})$	89.0	2.5	8.5	3.9	<0.1	14.5(0.1)
$\text{SVMIC}_H(L_n = \log(n))$	0.0	0.0	100.0	2.2	0.0	20.1(0.1)
$\text{SVMIC}_H(L_n = n^{-1/3})$	0.0	0.0	100.0	2.1	0.0	20.8(0.1)
Chen&Chen	74.0	0.5	25.5	3.7	<0.1	15.2(0.1)

performances of five-fold cross-validation and its adjusted version are slightly worse than those of SVMIC_H with L_n fixed at $\log(\log(n))$ and $\sqrt{\log(n)}$. Notice that the computation burden of selecting tuning parameter via information criterion is much lower than cross-validation. This makes our proposed information criteria desirable especially in the case with very large p.

5. Real data examples

5.1 MAQC-II breast cancer data

In this section we consider a real-world example from the breast cancer dataset which is part of the MicroArray Quality Control (MAQC)-II project. The preprocessed data can be downloaded from GEO databases with accession number GSE20194. There are 278 patient samples in the data and each is described by 22283 genes. Among the 278 samples, 164 patients have positive estrogen receptor (ER) status and 114 have negative ER status. Our goal is to predict the biological endpoint labeled by ER status and pick up the relevant genes.

We randomly choose 50 samples from positive ER status and 50 samples from negative ER status as the training data. The remaining 114 positive and 64 negatives are used for evaluating the prediction error, resulting in a test data of 178 patients. The data are

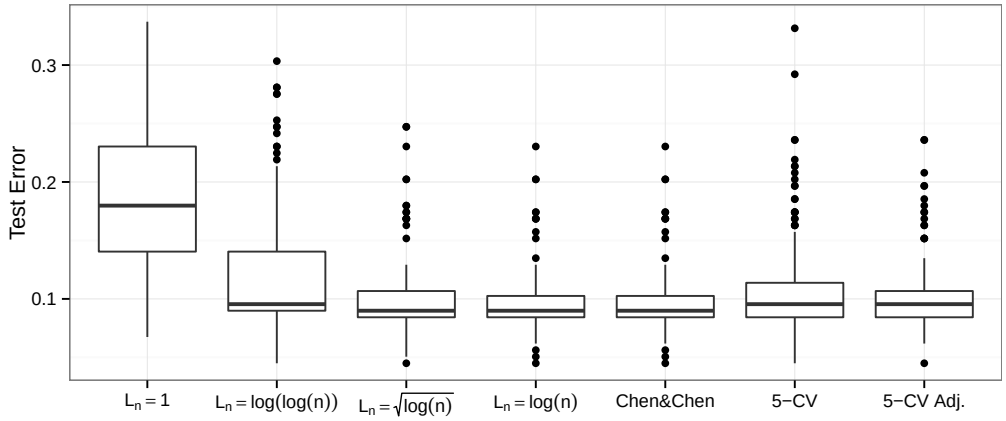
Table 3: Results for tuning parameter selection over 100 replications

Method	C(%)	O(%)	U(%)	Signal	Noise	Test Error(%)
		$n = 200,\ p = 2000$				
SVMIC_L	12	88	0	5.0	2.7	9.6(0.2)
$\text{SVMIC}_H(L_n = \log(\log(n)))$	78	22	0	5.0	0.3	7.4(0.1)
$\text{SVMIC}_H(L_n = \sqrt{\log(n)})$	97	3	0	5.0	<0.1	7.0(0.1)
$\text{SVMIC}_H(L_n = \log(n))$	64	0	36	4.3	0.0	11.5(0.7)
$\text{SVMIC}_H(L_n = n^{-1/3})$	58	0	42	4.1	0.0	12.5(0.8)
Chen&Chen	98	0	2	4.9	0.0	7.2(0.2)
5-CV	44	56	0	5.0	1.5	7.6(0.1)
5-CV Adj.	67	33	0	5.0	0.9	7.5(0.1)
		$n = 300,\ p = 3000$				
SVMIC_L	29	71	0	5.0	3.1	8.8(0.2)
$\text{SVMIC}_H(L_n = \log(\log n))$	94	6	0	5.0	0.1	6.9(0.1)
$\text{SVMIC}_H(L_n = \sqrt{\log(n)})$	100	0	0	5.0	0.0	6.8(0.1)
$\text{SVMIC}_H(L_n = \log(n))$	96	0	4	4.9	0.0	7.1(0.1)
$\text{SVMIC}_H(L_n = n^{-1/3})$	90	0	10	4.8	0.0	7.7(0.3)
Chen&Chen	100	0	0	5.0	0.0	6.8(0.1)
5-CV	58	42	0	5.0	1.0	7.2(0.1)
5-CV Adj.	76	24	0	5.0	0.7	7.1(0.1)

standardized before fitting the classifier. To reduce the computation burden, only 3000 genes with largest absolute values of the two sample t-statistics are used. Such simplification has been considered in Cai and Liu (2011). Though only 3000 genes are used, the classification result is satisfactory. We implement the SCAD penalized SVM to construct the solution path and set the range of λ as $\{2^{-15}, 2^{-14}, \ldots, 2^3\}$. The models on the solution path are selected by SVMIC_L (equivalent to $L_n = 1$), SVMIC_H with L_n at $\log(\log(n))$, $\sqrt{\log(n)}$ and $\log(n)$, the criterion adapted from Chen and Chen (2008), five-fold cross-validations and its adjusted version. This procedure is repeated for 200 replications. The corresponding standard errors are summarized in parentheses. Notice that the 3000 genes with the largest absolute values of t-statistics are pre-selected only using the training data to avoid overfitting so they may be different across the 200 random partitions of the data.

Table 4 summarizes the averages and standard errors for MAQC-II breast cancer data. The criterion $\text{SVMIC}_H(\lambda)$ performs uniformly better than $\text{SVMIC}_L(\lambda)$ regardless of the choice of L_n. It can be easily seen that $\text{SVMIC}_L(\lambda)$ leads to overfitted models in this dataset and has a significant higher misclassification rate. This is in accordance with the theoretical findings in Section 3 that SVMIC_L can be too liberal when the sample size is not comparable to the number of features, while SVMIC_H is a consistent model selection criterion. For this dataset, SVMIC_H with L_n at $\sqrt{\log(n)}$ and $\log(n)$ and the criterion from Chen and Chen (2008) perform the best and are slightly better than cross-validation methods. As in previous arguments, the criterion adapted from the EBIC in Chen and Chen (2008) performs similarly as SVMIC_H with L_n between $\log(\log(n))$ and $\log(n)$ for a wide range of combinations of n and p. Figure 1 summarizes the distributions of test errors over the 200 random partitions of the data for different methods. Note that SVMIC_H is a more

Figure 1: Test error for MAQC-II breast cancer datasets over 200 random partitions



stable method than cross-validations across the partitions of the data. Furthermore, cross-validation based on data resampling is more computationally intensive and this discrepancy is expected to increase dramatically if we take all the genes into consideration, which makes the cross-validation less feasible than information criterion method.

Table 4: Results for MAQC-II breast cancer datasets over 200 random partitions

Method	Size	Test Error(%)
SVMIC_L	4.0	18.5(0.4)
$\text{SVMIC}_H(L_n = \log(\log n))$	1.7	11.5(0.4)
$\text{SVMIC}_H(L_n = \sqrt{\log(n)})$	1.2	9.9(0.2)
$\text{SVMIC}_H(L_n = \log(n))$	1.1	9.6(0.2)
Chen&Chen	1.1	9.6(0.2)
5-CV	7.7	10.8(0.3)
5-CV Adj.	5.1	10.1(0.2)

6. Discussion

In this paper we consider model selection information criterion for support vector machines in the diverging model space. We show that the information criterion proposed in Claeskens et al. (2008) is consistent when the number of features is fixed but can be too liberal if the dimensionality is diverging. A new support vector machine information criterion is proposed for model selection in high dimensions. Based on the uniform convergence rate, we prove that the new information criterion enjoys the model selection consistency even when the number of variables diverges exponentially fast with the sample size. We also link this information criterion to tuning parameter selection for penalized support vector machines. The proposed information criterion is more scalable and easier to compute than resampling techniques such as cross-validation. Simulations and real data examples confirm the model

selection consistency and the ability of selecting tuning parameter when the number of features is much larger than the sample size.

There are several issues yet to be investigated. In this paper we assume that the size of the true model is fixed and the smallest signal does not diminish to zero as the sample size increases. Minimum signal condition has been used in many papers including Fan and Peng (2004) and Fan and Lv (2011). It seems that our condition is stronger than theirs. It is possible to relax this condition. We could possibly assume that $q = q_n$ diverges with n such that $q_n = O(n^{a_1})$ for some $0 \leq a_1 < 1/2$. Then we can allow the minimum magnitude of the nonzero-signal to diminish to zero at an appropriate rate such as $\min_{1\leq j\leq q_n} |\beta_j^*| > an^{-(1-a_2)/2}$ for some constant $a > 0$ and $2a_1 < a_2 \leq 1$. In general, the condition we impose on $\min_{1\leq j\leq q_n} |\beta_j^*|$ is intertwined with the conditions on q and the matrix $\mathbf{X}$, which would be the same for any other high-dimensional regression problem. For a detailed discussion on the beta-min condition in the setting of Lasso regression, we refer to Section 7.4 of Bühlmann et al. (2011). Another direction of interest is to extend the information criterion to nonlinear support vector machine. It is well known that the linear support vector machine can be easily extended to nonlinear feature space using the "kernel trick". Note that it is possible to extend the results in this paper to reproducing kernel Hilbert space with polynomial kernels. For Gaussian radial basis kernels, however, the direct generalization can be problematic as the corresponding reproducing kernel Hilbert space is infinite dimensional. A refined definition of the size of model will be needed in that case and will lead to a more comprehensive study of support vector machine information criterion.

Acknowledgments

We thank the Action Editor and two referees for very constructive comments and suggestions which have improved the presentation of the paper. Wu's research is partially supported by NSF grant DMS-1055210, NIH/NCI grants P01-CA142538 and R01-CA149569. Wang's research is supported by NSF grant DMS-1308960. Li's research was partially supported by NIH/NIDA grants P50-DA10075 and P50-DA036107. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NSF, NIH, NCI, or NIDA.

Appendix A.

In this appendix we prove the following results from Section 3.2:

Lemma 1 *Assuming p is a fixed number and $\lambda_n = 1/n$. Under conditions (A1)-(A4) and (A6)-(A7), we have*

$$\Pr(\widehat{S} = S^*) \to 1$$

as $n \to \infty$, where $\widehat{S} = \arg\min_{S:|S|\le M} SVMIC_L(S)$. ■

Proof. Under regularity conditions, Koo et al. (2008) showed $\widehat{\boldsymbol{\beta}}_S$ is root-n consistent in fixed p case for every $S \in \{S : |S| \le M_n\}$. This pointwise result is enough for Lemma 1 since the model space is fixed. The proof is then similar to Lemma 3 and Lemma 4 for diverging p with the uniform convergence rate $\sqrt{|S|\log(p)/n}$ substituted by $\sqrt{n}^{-1}$ and thus is omitted here.

Lemma 2 *Under conditions (A1)-(A7) and $\lambda_n = 1/n$, we have*

$$\sup_{S:|S|<M_n, S\supset S^*} ||\widehat{\boldsymbol{\beta}}_S - \boldsymbol{\beta}^*_S|| = O_p(\sqrt{|S|\log(p)/n}).$$

■

Proof. Recall that $\widehat{\boldsymbol{\beta}}_S = \arg\min_{\boldsymbol{\beta}_S}\{1/n\sum_{i=1}^n(1 - Y_i\mathbf{X}_{i,S}^T\boldsymbol{\beta}_S)_+ + \lambda_n/2||\boldsymbol{\beta}_S^+||^2\}$. We will show that for any $0 < \eta < 1$, there exists a large constant $\triangle$ such that for sufficient large n,

$$\Pr(\inf_{|S|\le M_n, S\supset S^*}\inf_{||\mathbf{u}||=\triangle} l_S(\boldsymbol{\beta}^*_S + \sqrt{|S|\log(p)/n}\mathbf{u}) > l_S(\boldsymbol{\beta}^*_S)) > 1 - \eta$$

where $l_S(\boldsymbol{\beta}_S) = 1/n\sum_{i=1}^n(1 - Y_i\mathbf{X}_{i,S}^T\boldsymbol{\beta}_S)_+ + \lambda_n/2||\boldsymbol{\beta}_S^+||^2$. By the convexity of the hinge loss, this implies that with probability $1-\eta$, we have $\sup_{S:|S|\le M_n, S\supset S^*}||\widehat{\boldsymbol{\beta}}_S - \boldsymbol{\beta}^*_S|| \le \triangle\sqrt{|S|\log(p)/n}$ and thus Lemma 2 holds.

Notice that we can decompose $l_S(\boldsymbol{\beta}^*_S + \sqrt{|S|\log(p)/n}\mathbf{u}) - l_S(\boldsymbol{\beta}^*_S)$ as

$$\begin{aligned}
&l_S(\boldsymbol{\beta}^*_S + \sqrt{|S|\log(p)/n}\mathbf{u}) - l_S(\boldsymbol{\beta}^*_S)\\
=&1/n\sum_{i=1}^n\{(1 - Y_i\mathbf{X}_{i,S}^T(\boldsymbol{\beta}^*_S + \sqrt{|S|\log(p)/n}\mathbf{u}))_+ - (1 - Y_i\mathbf{X}_{i,S}^T\boldsymbol{\beta}^*_S)_+\}\\
&+\lambda_n/2||\boldsymbol{\beta}^{*+}_S + \sqrt{|S|\log(p)/n}\mathbf{u}^+||^2 - \lambda_n/2||\boldsymbol{\beta}^{*+}_S||^2. \qquad (10)
\end{aligned}$$

By the fact $||\boldsymbol{\beta}^{*+}_S + \sqrt{|S|\log(p)/n}\mathbf{u}^+||^2 - ||\boldsymbol{\beta}^{*+}_S||^2 \le \triangle|S|\sqrt{\log(p)|S|/n}$ and $\lambda_n = 1/n$, the difference of penalty terms in (10) is $n^{-1}|S|o(1)$. Denote

$$\begin{aligned}
g_{i,S}(\mathbf{u}) =&(1 - Y_i\mathbf{X}_{i,S}^T(\boldsymbol{\beta}^*_S + \sqrt{|S|\log(p)/n}\mathbf{u}))_+ - (1 - Y_i\mathbf{X}_{i,S}^T\boldsymbol{\beta}^*_S)_+\\
&+\sqrt{|S|\log(p)/n}Y_i\mathbf{X}_{i,S}^T\mathbf{u}\mathbf{1}(1 - Y_i\mathbf{X}_{i,S}^T\boldsymbol{\beta}^*_S \ge 0)\\
&+\mathbb{E}[(1 - Y_i\mathbf{X}_{i,S}^T(\boldsymbol{\beta}^*_S + \sqrt{|S|\log(p)/n}\mathbf{u}))_+] - \mathbb{E}[(1 - Y_i\mathbf{X}_{i,S}^T\boldsymbol{\beta}^*_S)_+].
\end{aligned}$$

It can easily checked that $\mathbb{E}[g_{i,S}(\mathbf{u})] = 0$ for $\{S : |S| \le M_n, S \supset S^*\}$ by the definition of $\boldsymbol{\beta}^*_S$ and $\mathbf{S}(\boldsymbol{\beta}^*) = \mathbf{0}$. Next we consider the difference of hinge loss in (10), which can be further composed as

$$1/n\sum_{i=1}^n\{(1 - Y_i\mathbf{X}_{i,S}^T(\boldsymbol{\beta}^*_S + \sqrt{|S|\log(p)/n}\mathbf{u}))_+ - (1 - Y_i\mathbf{X}_{i,S}^T\boldsymbol{\beta}^*_S)_+\} = 1/n(A_n + B_n),$$

where

$$A_n = \sum_{i=1}^{n} g_{i,S}(\mathbf{u})$$

and

$$B_n = \sum_{i=1}^{n} \big\{ -\sqrt{|S|\log(p)/n} Y_i \mathbf{X}_{i,S}^T \mathbf{u}\mathbf{1}(1 - Y_i \mathbf{X}_{i,S}^T \boldsymbol{\beta}_S^* \geq 0) \\ + \mathbb{E}[(1 - Y_i \mathbf{X}_{i,S}^T (\boldsymbol{\beta}_S^* + \sqrt{|S|\log(p)/n}\mathbf{u}))_+] - \mathbb{E}[(1 - Y_i \mathbf{X}_{i,S}^T \boldsymbol{\beta}_S^*)_+] \big\}.$$

The rest of the proof consists of three steps. Step 1 will show

$$\sup_{|S| \leq M_n, S \supset S^*} \sup_{||\mathbf{u}||=\triangle} |A_n| = |S| o_p(1).$$

Step 2 will show $\inf_{|S| \leq M_n, S \supset S^*} \inf_{||\mathbf{u}||=\triangle} B_n$ dominates the terms of order $|S| o_p(1)$. Step 3 will complete the proof by showing $\inf_{|S| \leq M_n, S \supset S^*} \inf_{||\mathbf{u}||=\triangle} B_n > 0$ for sufficient large n and $\triangle$.

Step 1: The main tool to prove this uniform rate is the covering number introduced in Van Der Vaart and Wellner (1996). It suffices to show that

$$\Pr(\sup_{|S| \leq M_n, S \supset S^*} \sup_{||\mathbf{u}||=\triangle} |S|^{-1} | \sum_{i=1}^{n} g_{i,S}(\mathbf{u}|) > \epsilon) \to 0$$

for any $\epsilon > 0$. Notice that the hinge loss satisfies Lipschitz condition and by condition (A3) $\max_i ||\mathbf{X}_{i,S}|| = O_p(\sqrt{|S|}\log(n))$. It can be easily shown that

$$|S|^{-1} g_{i,S}(\mathbf{u}) \leq 3\triangle |S|^{-1} \sqrt{|S|\log(p)/n} \max_i ||\mathbf{X}_{i,S}||$$

and thus $\sup_{|S| \leq M_n, S \supset S^*} \sup_{||\mathbf{u}||=\triangle} |S|^{-1} g_{i,S}(\mathbf{u}) = o_p(1)$. By Lemma 2.5 of van de Geer (2000), the ball $\{\mathbf{u} : ||\mathbf{u}|| \leq \triangle\}$ in $\mathbf{R}^{|S|+1}$ can be covered by N balls with radius δ where $N \leq ((4\triangle + \delta)/\delta)^{|S|+1}$. Denote $\mathbf{u}^1, \ldots, \mathbf{u}^N$ the centers of the N balls. By the fact that $\sup_{|S| \leq M_n, S \supset S^*} \sqrt{|S|\log(p)/n} \max_i ||\mathbf{X}_{i,S}|| = O_p(1)$, we can take $\delta = (nC)^{-1}|S|$ for some large constant C such that

$$\min_{1 \leq k \leq N} \sup_{|S| \leq M_n, S \supset S^*} \sup_{||\mathbf{u}||=\triangle} |S|^{-1} | \sum_{i=1}^{n} g_{i,S}(\mathbf{u}) - \sum_{i=1}^{n} g_{i,S}(\mathbf{u}^k)| \\ \leq \sup_{|S| \leq M_n, S \supset S^*} 3\triangle n |S|^{-1} \sqrt{|S|\log(p)/n} \max_i ||\mathbf{X}_{i,S}|| \delta \ \leq \ \epsilon/3 \tag{11}$$

with probability tending to one. Based on (11), it can be easily shown

$$\Pr(\sup_{|S| \leq M_n, S \supset S^*} \sup_{||\mathbf{u}||=\triangle} |S|^{-1} | \sum_{i=1}^{n} g_{i,S}(\mathbf{u})| > \epsilon) \\ \leq \sum_{|S| \leq M_n, S \supset S^*} \sum_{k=1}^{N} \Pr(|S|^{-1} | \sum_{i=1}^{n} g_{i,S}(\mathbf{u}^k)| > \epsilon/2)$$

and $\sum_{i=1}^n g_{i,S}(\mathbf{u}^k)$ is sum of independent zero-mean random variables. Notice that

$$(1 - Y_i\mathbf{X}_{i,S}^T(\boldsymbol{\beta}_S^* + \sqrt{|S|\log(p)/n}\mathbf{u}))_+ - (1 - Y_i\mathbf{X}_{i,S}^T\boldsymbol{\beta}_S^*)_+ \\ +\sqrt{|S|\log(p)/n}Y_i\mathbf{X}_{i,S}^T\mathbf{u}\mathbf{1}(1 - Y_i\mathbf{X}_{i,S}^T\boldsymbol{\beta}_S^* \geq 0) = 0$$

when we have $|1 - Y_i\mathbf{X}_{i,S}^T\boldsymbol{\beta}_S^*| > \sqrt{|S|\log(p)/n}\max_i ||\mathbf{X}_{i,S}||\triangle$. Thus we have

$$\sum_{i=1}^n \mathbb{E}[g_{i,S}(\mathbf{u}^k)]^2 = \sum_{i=1}^n \mathrm{Var}(g_{i,S}(\mathbf{u}^k)) \\ \leq \sum_{i=1}^n \mathbb{E}[(2\sqrt{|S|\log(p)/n}Y_i\mathbf{X}_{i,S}^T\mathbf{u}^k)^2\mathbf{1}(|1 - Y_i\mathbf{X}_{i,S}^T\boldsymbol{\beta}_S^*| \leq \sqrt{|S|\log(p)/n}\max_i ||\mathbf{X}_{i,S}||\triangle)]. \tag{12}$$

By the bounded largest eigenvalue condition in (A3), we have

$$\sum_{i=1}^n \mathbb{E}\{[2\sqrt{|S|\log(p)/n}Y_i\mathbf{X}_{i,S}^T\mathbf{u}^k)]^2 \leq C|S|\log(p).$$

By the bounded conditional density condition (A4), we have

$$\Pr(|1 - Y_i\mathbf{X}_{i,S}^T\boldsymbol{\beta}_S^*| \leq \sqrt{|S|\log(p)/n}\max_i ||\mathbf{X}_{i,S}||\triangle) \leq C|S|\log n\sqrt{\log(p)/n}.$$

Then based on (12) and Cauchy inequality, we have

$$\sum_{i=1}^n \mathbb{E}[g_{i,S}(\mathbf{u}^k)]^2 \leq C|S|^2\log n(\log(p))^{3/2}n^{-1/2}.$$

Then applying Bernstein inequality and condition (A6), we arrive

$$\sum_{|S|\leq M_n, S\supset S^*}\sum_{k=1}^N \Pr(|S|^{-1}|\sum_{i=1}^n g_{i,S}(\mathbf{u}^k)| > \epsilon/2) \\ \leq \exp\{M_n\log(p)\}\exp(N)\exp\{-C(\log(n))^{-1}(\log(p))^{-3/2}n^{1/2}\} \to 0$$

as $n \to \infty$. This completes the proof of Step 1.

Step 2: First notice that

$$|\sum_{i=1}^n Y_i\mathbf{X}_{i,S}^T\mathbf{u}\mathbf{1}(1 - Y_i\mathbf{X}_{i,S}^T\boldsymbol{\beta}_S^* \geq 0)| \leq (|S|+1)^{1/2}\triangle \max_{0\leq j\leq p}|\sum_{i=1}^n Y_iX_{ij,S}\mathbf{1}(1 - Y_i\mathbf{X}_{i,S}^T\boldsymbol{\beta}_S^* \geq 0)|. \tag{13}$$

Note that $\mathbb{E}[Y_iX_{ij,S}\mathbf{1}(1 - Y_i\mathbf{X}_{i,S}^T\boldsymbol{\beta}_S^* \geq 0)] = 0$ for $0 \leq j \leq p$ by the definition of $\mathbf{S}(\boldsymbol{\beta}^*)$. By Lemma 14.24 of Bühlmann et al. (2011), we also have

$$\max_{0\leq j\leq p}|\sum_{i=1}^n Y_iX_{ij,S}\mathbf{1}(1 - Y_i\mathbf{X}_{i,S}^T\boldsymbol{\beta}_S^* \geq 0)| = O_p(\sqrt{n\log(p)}). \tag{14}$$

By Taylor expansion of hinge loss function at $\boldsymbol{\beta}_S^*$, we have

$$\begin{aligned}&\sum_{i=1}^{n}\{\mathbb{E}[(1-Y_i\mathbf{X}_{i,S}^T(\boldsymbol{\beta}_S^*+\sqrt{|S|\log(p)/n}\mathbf{u}))_+]-\mathbb{E}[(1-Y_i\mathbf{X}_{i,S}^T\boldsymbol{\beta}_S^*)_+]\}\\ =&0.5|S|\log(p)\mathbf{u}^T\mathbf{H}(\boldsymbol{\beta}_S^*+t\sqrt{|S|\log(p)/n}\mathbf{u})\mathbf{u} \qquad (15)\end{aligned}$$

for some $0<t<1$. As shown by Koo et al. (2008), under condition (A1) and (A2), $\mathbf{H}(\boldsymbol{\beta})$ is element-wise continuous at $\boldsymbol{\beta}_S^*$, thus

$$\mathbf{H}(\boldsymbol{\beta}_S^*+t\sqrt{|S|\log p/n}\mathbf{u})=\mathbf{H}(\boldsymbol{\beta}_S^*)+o_p(1).$$

It can be easily shown by (13), (14), (15) and condition (A7), $0.5|S|\log(p)\mathbf{u}^T\mathbf{H}(\boldsymbol{\beta}_S^*)\mathbf{u}$ dominates other terms in B_n for sufficient large $\triangle$. This completes the proof of Step 2.

Step 3: Notice that $0.5|S|\log(p)\mathbf{u}^T\mathbf{H}(\boldsymbol{\beta}_S^*)\mathbf{u}>0$ by condition (A7). Recall that the difference of penalty terms in (10) is $n^{-1}|S|o(1)$. Therefore $0.5|S|\log p\mathbf{u}^T\mathbf{H}(\boldsymbol{\beta}_S^*)\mathbf{u}$ dominates all the other terms in (10) for sufficient large n and $\triangle$, which completes the proof.

Lemma 3 *Under conditions (A1)-(A8) and $\lambda_n=1/n$, we have*

$$\Pr(\inf_{S:S\in\Omega_+} SVMIC_H(S)>SVMIC_H(S^*))\to 1.$$

as $n\to\infty$. ■

Proof. By definition we have

$$\begin{aligned}&\inf_{S\in\Omega_+}\mathrm{SVMIC}_H(S)-\mathrm{SVMIC}_H(S^*)\\ =&\inf_{S\in\Omega_+}\big\{\sum_{i=1}^{n}(1-Y_i\mathbf{X}_{i,S}^T\widehat{\boldsymbol{\beta}}_S)_+-\sum_{i=1}^{n}(1-Y_i\mathbf{X}_{i,S^*}^T\widehat{\boldsymbol{\beta}}_{S^*})_++(|S|-|S^*|)\log(n)L_n\big\}.\end{aligned}$$

Similar to the proof of Lemma 2, it can be shown that $|\sum_{i=1}^{n}(1-Y_i\mathbf{X}_{i,S}^T\widehat{\boldsymbol{\beta}}_S)_+-\sum_{i=1}^{n}(1-Y_i\mathbf{X}_{i,S^*}^T\widehat{\boldsymbol{\beta}}_{S^*})_+|$ is dominated by $|S|\log(p)\mathbf{u}^T\mathbf{H}(\boldsymbol{\beta}_S^*)\mathbf{u}$ with probability tending to one. By conditions (A6)-(A8), we have

$$|\sum_{i=1}^{n}(1-Y_i\mathbf{X}_{i,S}^T\widehat{\boldsymbol{\beta}}_S)_+-\sum_{i=1}^{n}(1-Y_i\mathbf{X}_{i,S^*}^T\widehat{\boldsymbol{\beta}}_{S^*})_+|<(|S|-|S^*|)\log(n)L_n$$

for sufficient large n. Notice that $\inf_{S\in\Omega_+}|S|-|S^*|>0$, which completes the proof.

Lemma 4 *Under Conditions (A1)-(A8) and $\lambda_n=1/n$, we have*

$$\Pr(\inf_{S:S\in\Omega_-} SVMIC_H(S)>SVMIC_H(S^*))\to 1.$$

as $n\to\infty$. ■

Proof. For $S\in\Omega_-$, consider the set $\tilde{S}$ with additional signals such that $\tilde{S}=S\cup S^*$. Notice

$$\mathrm{SVMIC}_H(S)-\mathrm{SVMIC}_H(S^*)=\mathrm{SVMIC}_H(S)-\mathrm{SVMIC}_H(\tilde{S})+\mathrm{SVMIC}_H(\tilde{S})-\mathrm{SVMIC}_H(S^*)$$

for $S \in \Omega_-$. Since $|S^*|$ does not diverge with n, we have $|\tilde{S}| < 2M_n$ for sufficiently large n and it can easily seen that Lemma 3 still holds for $\tilde{S}$ with any $S \in \Omega_-$. Therefore with high probability we have $\text{SVMIC}_H(\tilde{S}) - \text{SVMIC}_H(S^*) \geq 0$. Thus it suffices to show

$$\Pr(\inf_{S\in\Omega_-}\{\text{SVMIC}_H(S) - \text{SVMIC}_H(\tilde{S})\} > 0) \to 1$$

as $n \to \infty$. Notice that

$$\begin{aligned}
&1/n\{\text{SVMIC}_H(S) - \text{SVMIC}_H(\tilde{S})\} \\
=&1/n\sum_{i=1}^{n}(1 - Y_i\mathbf{X}_{i,S}^T\widehat{\boldsymbol{\beta}}_S)_+ - 1/n\sum_{i=1}^{n}(1 - Y_i\mathbf{X}_{i,\tilde{S}}^T\widehat{\boldsymbol{\beta}}_{\tilde{S}})_+ + 1/n(|S| - |\tilde{S}|)\log(n)L_n
\end{aligned}$$

and by condition (A8) $1/n(|S| - |\tilde{S}|)\log(n)L_n \to 0$, it suffices to show

$$\inf_{S\in\Omega_-}\{1/n\sum_{i=1}^{n}(1 - Y_i\mathbf{X}_{i,S}^T\widehat{\boldsymbol{\beta}}_S)_+ - 1/n\sum_{i=1}^{n}(1 - Y_i\mathbf{X}_{i,\tilde{S}}^T\widehat{\boldsymbol{\beta}}_{\tilde{S}})_+\} \geq C$$

for some constant $C > 0$ that does not depend on S.

Recall that $\widehat{\boldsymbol{\beta}}_S = (\beta_{0,S}, \beta_{1,S}, \ldots, \beta_{|S|,S})^T \in \mathbf{R}^{|S|+1}$. Denote $\widehat{\boldsymbol{\beta}}_{S,\tilde{S}} \in \mathbf{R}^{|\tilde{S}|+1}$ such that the intercept equals to $\beta_{0,S}$, the j-th element equals to $\beta_{j,S}$ if $j \in S$ and 0 if $j \notin S$ for all $j \in \tilde{S}$. Denote also $\delta = \min_{j\in S^*}|\beta_j^*|$ the smallest signal. Then it can be easily seen that $||\widehat{\boldsymbol{\beta}}_{S,\tilde{S}} - \boldsymbol{\beta}_{\tilde{S}}^*|| > \delta$. By Lemma 2 we also have $||\widehat{\boldsymbol{\beta}}_{\tilde{S}} - \boldsymbol{\beta}_{\tilde{S}}^*|| < \epsilon$ for arbitrary ϵ and sufficient large n. Therefore there exists $\bar{\boldsymbol{\beta}}_{\tilde{S}} = a\widehat{\boldsymbol{\beta}}_{\tilde{S}} + (1-a)\widehat{\boldsymbol{\beta}}_{S,\tilde{S}}$ for some $0 < a < 1$ such that

$$||\bar{\boldsymbol{\beta}}_{\tilde{S}} - \boldsymbol{\beta}_{\tilde{S}}^*|| = \Delta,$$

where Δ is a positive constant such that $\Delta < c_3$ where c_3 is defined in condition (A7). By the definition of $\widehat{\boldsymbol{\beta}}_{\tilde{S}}$ and the convexity of hinge loss function we have

$$\begin{aligned}
&1/n\sum_{i=1}^{n}(1 - Y_i\mathbf{X}_{i,\tilde{S}}^T\bar{\boldsymbol{\beta}}_{\tilde{S}})_+ + \lambda_n/2||\bar{\boldsymbol{\beta}}_{\tilde{S}}^+||^2 \\
<&a\{1/n\sum_{i=1}^{n}(1 - Y_i\mathbf{X}_{i,\tilde{S}}^T\widehat{\boldsymbol{\beta}}_{\tilde{S}})_+ + \lambda_n/2||\widehat{\boldsymbol{\beta}}_{\tilde{S}}^+||^2\} \\
&+ (1-a)\{1/n\sum_{i=1}^{n}(1 - Y_i\mathbf{X}_{i,S}^T\widehat{\boldsymbol{\beta}}_{S,\tilde{S}})_+ + \lambda_n/2||\widehat{\boldsymbol{\beta}}_{S,\tilde{S}}^+||^2\} \\
<&1/n\sum_{i=1}^{n}(1 - Y_i\mathbf{X}_{i,\tilde{S}}^T\widehat{\boldsymbol{\beta}}_{S,\tilde{S}})_+ + \lambda_n/2||\widehat{\boldsymbol{\beta}}_{S,\tilde{S}}^+||^2 \\
=&1/n\sum_{i=1}^{n}(1 - Y_i\mathbf{X}_{i,S}^T\widehat{\boldsymbol{\beta}}_S)_+ + \lambda_n/2||\widehat{\boldsymbol{\beta}}_S^+||^2. \qquad (16)
\end{aligned}$$

By $\lambda_n = n^{-1}$ we have

$$\lambda_n/2||\bar{\boldsymbol{\beta}}_{\tilde{S}}^+||^2 - \lambda_n/2||\widehat{\boldsymbol{\beta}}_S^+||^2 \leq C\lambda_n(||\widehat{\boldsymbol{\beta}}_{\tilde{S}}^+||^2 + ||\widehat{\boldsymbol{\beta}}_S^+||^2) \to 0 \qquad (17)$$

as $n \to \infty$. Similar to the proof of Lemma 2, under condition (A6) and (A8), it can be shown

$$1/n\sum_{i=1}^{n}(1-Y_i\mathbf{X}_{i,\tilde{S}}^T\widehat{\boldsymbol{\beta}}_{\tilde{S}})_+ - 1/n\sum_{i=1}^{n}(1-Y_i\mathbf{X}_{i,\tilde{S}}^T\boldsymbol{\beta}_{\tilde{S}}^*)_+ \le 1/nC|\tilde{S}|\log(p)\lambda_{\max}(\mathbf{H}(\boldsymbol{\beta}_{\tilde{S}}^*)) \to 0 \quad (18)$$

as $n \to \infty$. Notice that

$$\begin{aligned}
&\inf_{S\in\Omega_-}\{1/n\sum_{i=1}^{n}(1-Y_i\mathbf{X}_{i,\tilde{S}}^T\bar{\boldsymbol{\beta}}_{\tilde{S}})_+ - 1/n\sum_{i=1}^{n}(1-Y_i\mathbf{X}_{i,\tilde{S}}^T\boldsymbol{\beta}_{\tilde{S}}^*)_+\}\\
\ge& 1/n\Big\{\inf_{S\in\Omega_-} n\mathbb{E}[(1-Y_i\mathbf{X}_{i,\tilde{S}}^T\bar{\boldsymbol{\beta}}_{\tilde{S}})_+ - (1-Y_i\mathbf{X}_{i,\tilde{S}}^T\boldsymbol{\beta}_{\tilde{S}}^*)_+]\\
&-\sup_{S\in\Omega_-}\{|\sum_{i=1}^{n}(1-Y_i\mathbf{X}_{i,\tilde{S}}^T\bar{\boldsymbol{\beta}}_{\tilde{S}})_+ - \sum_{i=1}^{n}(1-Y_i\mathbf{X}_{i,\tilde{S}}^T\boldsymbol{\beta}_{\tilde{S}}^*)_+ - n\mathbb{E}[(1-Y_i\mathbf{X}_{i,\tilde{S}}^T\bar{\boldsymbol{\beta}}_{\tilde{S}})_+ - (1-Y_i\mathbf{X}_{i,\tilde{S}}^T\boldsymbol{\beta}_{\tilde{S}}^*)_+]|\}\Big\}.
\end{aligned}$$

Similar to the proof of Lemma 2, it can be shown

$$\begin{aligned}
&\sup_{S\in\Omega_-}\{|\sum_{i=1}^{n}(1-Y_i\mathbf{X}_{i,\tilde{S}}^T\bar{\boldsymbol{\beta}}_{\tilde{S}})_+ - \sum_{i=1}^{n}(1-Y_i\mathbf{X}_{i,\tilde{S}}^T\boldsymbol{\beta}_{\tilde{S}}^*)_+ - n\mathbb{E}[(1-Y_i\mathbf{X}_{i,\tilde{S}}^T\bar{\boldsymbol{\beta}}_{\tilde{S}})_+ - (1-Y_i\mathbf{X}_{i,\tilde{S}}^T\boldsymbol{\beta}_{\tilde{S}}^*)_+]|\}\\
=&O_p(|\sum_{i=1}^{n}Y_i\mathbf{X}_{i,\tilde{S}}^T(\bar{\boldsymbol{\beta}}_{\tilde{S}}-\boldsymbol{\beta}_{\tilde{S}}^*)\mathbf{1}(1-Y_i\mathbf{X}_{i,S}^T\boldsymbol{\beta}_{\tilde{S}}^*\ge 0)|) = O_p(\sqrt{n|\tilde{S}|\log(p)}). \qquad (19)
\end{aligned}$$

By Taylor expansion of hinge loss function, we have

$$\mathbb{E}[(1-Y_i\mathbf{X}_{i,\tilde{S}}^T\bar{\boldsymbol{\beta}}_{\tilde{S}})_+ - (1-Y_i\mathbf{X}_{i,\tilde{S}}^T\boldsymbol{\beta}_{\tilde{S}}^*)_+] \ge 0.5\lambda_{\min}(\mathbf{H}(\tilde{\boldsymbol{\beta}}_{\tilde{S}}^*))\Delta^2 > 0, \qquad (20)$$

where $\tilde{\boldsymbol{\beta}}_{\tilde{S}}^*$ lies in the set defined in condition (A7). By (16)-(20), we have

$$\inf_{S\in\Omega_-}\{1/n\sum_{i=1}^{n}(1-Y_i\mathbf{X}_{i,S}^T\widehat{\boldsymbol{\beta}}_S)_+ - 1/n\sum_{i=1}^{n}(1-Y_i\mathbf{X}_{i,\tilde{S}}^T\widehat{\boldsymbol{\beta}}_{\tilde{S}})_+\} \ge 0.5\lambda_{\min}(\mathbf{H}(\boldsymbol{\beta}_{\tilde{S}}^*))\Delta^2 > 0$$

for sufficient large n, which completes the proof.

Theorem 5 *Under conditions (A1)-(A8) and $\lambda_n = 1/n$, we have*

$$\Pr(\hat{S} = S^*) \to 1.$$

as n, $p \to \infty$, where $\widehat{S} = \arg\min_{S:|S|\le M_n} SVMIC_H(S)$. ■

Proof. The proof can be easily checked by combing the results from Lemma 3 and Lemma 4 and thus is omitted here.

References

Hirotugu Akaike. Information theory and an extension of the maximum likelihood principle. In *Second international symposium on information theory*, pages 267–281. Akademinai Kiado, 1973.

Natalia Becker, Grischa Toedt, Peter Lichter, and Axel Benner. Elastic scad as a novel penalization method for svm classification tasks in high-dimensional data. *BMC bioinformatics*, 12(1):138, 2011.

Paul S Bradley and Olvi L Mangasarian. Feature selection via concave minimization and support vector machines. In *ICML*, volume 98, pages 82–90, 1998.

Peter Lukas Bühlmann, Sara A van de Geer, and Sara Van de Geer. *Statistics for high-dimensional data*. Springer, 2011.

Tony Cai and Weidong Liu. A direct estimation approach to sparse linear discriminant analysis. *Journal of the American Statistical Association*, 106(496), 2011.

Jiahua Chen and Zehua Chen. Extended bayesian information criteria for model selection with large model spaces. *Biometrika*, 95(3):759–771, 2008.

Gerda Claeskens, Christophe Croux, and Johan Van Kerckhoven. An information criterion for variable selection in support vector machines. *The Journal of Machine Learning Research*, 9:541–558, 2008.

J. Fan and J. Lv. Non-concave penalized likelihood with np-dimensionality. *IEEE Transactions on Information Theory*, 57:5467–5484, 2011.

J. Fan and H. Peng. On non-concave penalized likelihood with diverging number of parameters. *The Annals of Statistics*, 32:928–961, 2004.

Jianqing Fan and Yingying Fan. High dimensional classification using features annealed independence rules. *Annals of statistics*, 36(6):2605, 2008.

Jianqing Fan and Runze Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96(456):1348–1360, 2001.

Yingying Fan and Cheng Yong Tang. Tuning parameter selection in high dimensional penalized likelihood. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 75(3):531–552, 2013.

Isabelle Guyon and André Elisseeff. An introduction to variable and feature selection. *The Journal of Machine Learning Research*, 3:1157–1182, 2003.

Isabelle Guyon, Jason Weston, Stephen Barnhill, and Vladimir Vapnik. Gene selection for cancer classification using support vector machines. *Machine learning*, 46(1-3):389–422, 2002.

Trevor. Hastie, Robert. Tibshirani, and J Jerome H Friedman. *The elements of statistical learning*, volume 1. Springer New York, 2001.

Trevor Hastie, Saharon Rosset, Robert Tibshirani, and Ji Zhu. The entire regularization path for the support vector machine. In *Journal of Machine Learning Research*, pages 1391–1415, 2004.

Shuichi Kawano. Selection of tuning parameters in bridge regression models via bayesian information criterion. *Statistical Papers*, pages 1–17, 2012.

Ja-Yong Koo, Yoonkyung Lee, Yuwon Kim, and Changyi Park. A bahadur representation of the linear support vector machine. *The Journal of Machine Learning Research*, 9: 1343–1368, 2008.

Eun Ryung Lee, Hohsuk Noh, and Byeong U Park. Model selection via bayesian information criterion for quantile regression models. *Journal of the American Statistical Association*, 109(505):216–229, 2014.

Rahul Mazumder, Jerome H Friedman, and Trevor Hastie. Sparsenet: Coordinate descent with nonconvex penalties. *Journal of the American Statistical Association*, 106(495), 2011.

Gideon Schwarz. Estimating the dimension of a model. *The annals of statistics*, 6(2): 461–464, 1978.

Jun Shao. An asymptotic theory for linear model selection. *Statistica Sinica*, 7(2):221–242, 1997.

Peide Shi and Chih-Ling Tsai. Regression model selectiona residual likelihood approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(2):237–252, 2002.

Sara van de Geer. Empirical processes in m-estimation. cambridge series in statistical and probabilistic mathematics, 2000.

Aad W Van Der Vaart and Jon A Wellner. *Weak Convergence*. Springer, 1996.

Grace Wahba et al. Support vector machines, reproducing kernel hilbert spaces and the randomized gacv. *Advances in Kernel Methods-Support Vector Learning*, 6:69–87, 1999.

Hansheng Wang, Runze Li, and Chih-Ling Tsai. Tuning parameter selectors for the smoothly clipped absolute deviation method. *Biometrika*, 94(3):553–568, 2007.

Hansheng Wang, Bo Li, and Chenlei Leng. Shrinkage tuning parameter selection with a diverging number of parameters. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(3):671–683, 2009.

Lan Wang, Yichao Wu, and Runze Li. Quantile regression for analyzing heterogeneity in ultra-high dimension. *Journal of the American Statistical Association*, 107(497):214–222, 2012.

Li Wang, Ji Zhu, and Hui Zou. The doubly regularized support vector machine. *Statistica Sinica*, 16(2):589, 2006.

Marten Wegkamp and Ming Yuan. Support vector machines with a reject option. *Bernoulli*, 17:1368–1385, 2011.

Jason Weston, Sayan Mukherjee, Olivier Chapelle, Massimiliano Pontil, Tomaso Poggio, and Vladimir Vapnik. Feature selection for svms. In *NIPS*, volume 12, pages 668–674, 2000.

Ming Yuan. High dimensional inverse covariance matrix estimation via linear programming. *The Journal of Machine Learning Research*, 99:2261–2286, 2010.

Cun-Hui Zhang. Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics*, 38(2):894–942, 2010.

Cun-Hui Zhang and Jian Huang. The sparsity and bias of the lasso selection in high-dimensional linear regression. *The Annals of Statistics*, 36(4):1567–1594, 2008.

Hao Helen Zhang, Jeongyoun Ahn, Xiaodong Lin, and Cheolwoo Park. Gene selection using support vector machines with non-convex penalty. *Bioinformatics*, 22(1):88–95, 2006.

Xiang Zhang, Yichao Wu, Lan Wang, and Runze Li. Variable selection for support vector machines in moderately high dimensions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 2014.

Ji Zhu, Saharon Rosset, Trevor Hastie, and Rob Tibshirani. 1-norm support vector machines. *Advances in neural information processing systems*, 16(1):49–56, 2004.

Hui Zou and Runze Li. One-step sparse estimates in nonconcave penalized likelihood models. *Annals of statistics*, 36(4):1509, 2008.

Hui Zou and Ming Yuan. The fo-norm support vector machine. *Statistica Sinica*, 18: 379–398, 2008.